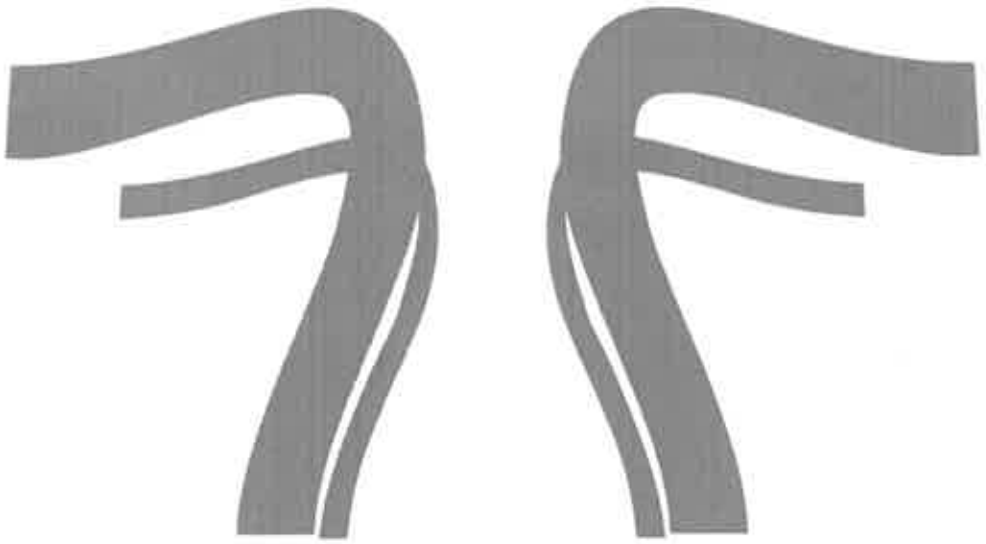
Proceedings of the 10th International Seminar on Speech Production (ISSP)

5 – 8 May 2014
Cologne, Germany



www.issp2014.uni-koeln.de

Edited by

Susanne Fuchs, Martine Grice, Anne Hermes, Leonardo Lancia, Doris Mücke

Measuring Reaction Times: Vocalisation vs. Articulation

Sonja Schaeffler, James M. Scobbie, Felix Schaeffler

CASL, Queen Margaret University, Edinburgh, UK

sschaeffler@qmu.ac.uk, jscobbie@qmu.ac.uk, fschaeffler@qmu.ac.uk

Abstract

There is a sizeable delay between any formulation of an intention to speak and the audible vocalisation that results. Silent articulatory movements in preparation for audible speech comprise a proportion of this phase of speech production. The extensive literature on Reaction Time (RT) is based on the delay between a stimulus and the acoustic onset to speech that is elicited, ignoring the preceding silent elements of speech production in what is an utterance-initial position. We used a standard Snodgrass and Vanderwart picture-naming task to elicit speech in a standard Reaction Time protocol, but recorded the behaviour of two typical speakers with audio plus Ultrasound Tongue Imaging (201 frames per second) and de-interlaced NTSC video of the mouth and lips (60fps). On average, Acoustic Reaction Time occurred between 120 to 180 ms later than a clearly observable articulatory movement, with no consistent advantage for lip or tongue-based measures.

Keywords: *Ultrasound Tongue Imaging, video, tongue, lips, Reaction Time, speech preparation*

1. Introduction

Like all organisms, humans have to react to a wide range of external stimuli in an appropriate way, within an appropriate time. A huge range of scientific studies related to such abilities have been undertaken, many specifically looking at Reaction Time, which is operationalised as the time it takes from presentation of a stimulus to an observable response. A very common observable response in psycholinguistic experiments is verbal feedback, and here, Reaction Time is usually determined via "voice key", a device which is triggered automatically as soon as the sound pressure reaches a pre-defined level. However, what is detectable as the onset of acoustic output is only one stage in the speech production process. Before anything becomes audible, the articulators (i.e. the tongue, the lips, the jaw, etc.) have already moved into place. This movement shows that the motor plan for the response has been put into action and is an earlier observable response than audible speech.

Models of speech production such as WEAVER++ (Roelofs 1997; Levelt et al 1999) or GODIVA (Bohland et al. 2012) are readily applied to single word productions, though as Mooshammer et al. (2012) point out, the models lack a great deal of detail and *"apart from the effects of learning/practice very little is understood about the assembly of motor (gestural) plans for an utterance"* (p. 375). We can assume in a single word Reaction Time task, the necessary sequential stages include extraction of phonologically encoded items from the lexicon, followed by the construction of a speech motor-plan for execution. The measurable acoustic response follows in turn, but silent movements of the articulators themselves are also clearly detectable by human observers, preceding audible output, and these can be quantified, given the appropriate instrumentation. (We should note that for longer phrases in connected speech, these processes overlap or occur in parallel, which is a more complex problem that need not concern us here.) Some instrumentation such as EMA has been used previously (Mooshammer et al. 2012) to look at this silent phase in production, though not directly to measure Reaction Time, but EMA is relatively expensive to use and is not generally available. Other instruments such as high-speed video or Ultrasound Tongue Imaging are much more readily available and with appropriate automated analysis methods (cf. Palo et al. 2014) could be a competitor to audio recordings plus voice key.

This paper is a pilot study aiming to investigate how much of a discrepancy there is between acoustic and articulatory onset of speech across a variety of tokens with different phonemic targets in a standard Reaction Time task. We obtained high-quality acoustic recordings to determine when verbal response becomes first audible, and used ultrasound recordings of the tongue, and video recordings of the lips, to observe when verbal response is first initiated by the articulators.

2. Method

2.1. Participants

We report data from two female native speakers of Scottish varieties of English (aged between between 20 and 35). The participants reported no visual or hearing impairments.

2.2. Material

Verbal responses were collected in a standard picture naming task, with target items drawn from the well-tested and frequently used Snodgrass-and-Vanderwart picture inventory (1980). The picture inventory consists of 260 targets.

2.3. Tongue and lip recordings: instrumentation

All articulatory and acoustic recordings were obtained simultaneously and synchronised using AAA software from Articulate Instruments Ltd (2012). The participants were fitted with a purpose-built headset to ensure stabilisation of the ultrasound probe (Articulate Instruments Ltd 2008; Scobbie et al. 2008). Attached to the helmet was a small Audio Technica AT803b microphone for high-quality acoustic recordings, plus a NTSC micro-camera to capture recordings of the speakers' lips.

Ultrasound recordings were obtained at a rate of 201 frames per second from a SonixRP system. Video was captured then deinterlaced to an effective rate of 59.95 fps. Recordings started 1.5 seconds before prompt presentation so that the whole speech production process was captured.

2.4. Synchronisation

Tests of synchronisation of audio to ultrasound, and of audio to video were made that suggest a <5ms error. Our audio-

ultrasound synchronisation uses a square-wave sync signal and each scan is stored as a unique frame with no overlap. Synchronisation at 201 fps is therefore within a 5 ms window. The video capture of the mouth and lips was synchronised using an Articulate Instruments 'bright-up' electronic clapperboard that records a sync pulse on camera and microphone outputs. Video frame rate was set at 19.97 fps. For both image types, we assume no extra delay in data transfer, or due to image creation (in contrast to video ultrasound, cf. Wrench et al. 2006).

3. Token selection, annotation and analysis

3.1. Tokens for analysis

Of the 260 targets presented, 19 elicited erroneous responses in Speaker 1 and were excluded, leaving 241 tokens for analysis. Speaker 2 could not tolerate the headset for the entire length of the experiment and stopped after recording of 156 tokens. Of the recorded tokens, 4 elicited erroneous responses, leaving 152 tokens for analysis for Speaker 2. Targets that elicited British English variants instead of the American English target (e.g. 'waistcoat' instead of 'vest') were not excluded.

3.2. Annotations

3.2.1. Labels

For each of the tokens (241 for Speaker 1; 152 for Speaker 2) we annotated a) the onset of target-related lip movement b) the onset of target-related tongue movement and c) the acoustic onset.

3.2.2. Articulatory annotation criteria

To be target-related, the relevant articulator had to be seen to move in a smooth and consistent manner towards one of the initial articulatory targets. For example a word like "barn" requires initial labial closure, and this might be detectable from an open mouth position or from a closed mouth that opens, then closes again for the /b/. We annotated only the unambiguous lip movement towards closure in each case. In addition, the tongue has to move to the appropriate position for the /a/. This might be from a neutral schwa-like position or from palatal contact. Again, only a single, contiguous unambiguous holistic tongue movement towards an articulatory target was annotated.

3.2.3. Acoustic annotation criteria

The acoustic onset was determined by visual inspection of the speech signal for a rapid and unambiguous increase in acoustic intensity and/or spectral cues to the onset of vocalisation. Figure 1 illustrates the different sources of data available for annotation in AAA.

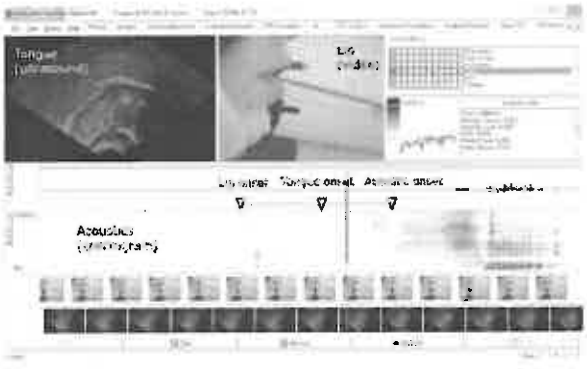


Figure 1: *Screenshot of AAA analysis window for Speaker 1 producing 'sock' with lips and tongue in place long before acoustic onset.*

3.3. Annotation confidence ratings

Given our hand-labelled holistic approach, we gave each articulatory annotation a confidence rating from '1 = very unsure' to '5 = absolutely sure'. Tokens that received a low rating included e.g. instances where it was difficult to determine the boundary between onset of target-related articulation and some directly preceding, non-linguistic pre-speech behaviour (e.g. tongue clicks, pursing of lips). For the purpose of exploring the suitability of our newly developed method for obtaining Articulatory Reaction Time data, we only used tokens for further analysis with annotations that for both lip and tongue onset received either a '4' or '5' confidence rating. This left us with 132 tokens (55%) for Speaker 1 and 82 tokens for Speaker 2 (54%).

3.4. Phonetic properties of targets

All tokens were coded in analysis for the phonetic properties of the actually uttered target word, i.e. 'Onset Voicing' (voiced / voiceless), 'Onset Manner' (vowel or glide / stop or affricate / fricative), 'Onset Place' (labial / lingual) and 'Number of Syllables' (monosyllabic / polysyllabic).

It is important to note that the tokens that had received lower confidence ratings of '1', '2' or '3' did not fall into any specific phonetic categories. An analysis of Speaker 1's excluded 109 tokens comprised of 56 stops/affricates, 36 fricatives and 17 vowels/glides with all places of articulation represented, compared to 67 stops/affricates, 35 fricatives and 30 vowels/glides in the confidently rated 132 tokens. This suggests that there was no place or manner of articulation that was inherently more difficult to determine articulatory onsets for.

4. Results

4.1. Acoustic Reaction Times

Across all remaining tokens the untrimmed Acoustic Reaction Time was on average 897 ms (SD 289 ms) for Speaker 1 and 624 ms (SD 205 ms) for Speaker 2. To exclude outliers that were caused by word finding difficulties or measurement artefacts we trimmed the sample by excluding all tokens with Acoustic Reaction Times that were higher or lower than the mean ±2SD (Speaker 1: n = 7; Speaker 2: n = 5). After removal of outliers the mean Acoustic Reaction Time was 851 ms (SD 251 ms) for Speaker 1 and 586 ms (SD 127 ms) for Speaker 2. That means overall Speaker 2 exhibited a much faster Reaction Time than Speaker 1.

4.2. Articulatory Reaction Times

Lip Reaction Time was on average 677 ms (SD 237 ms) for Speaker 1 and 466 ms (SD 127 ms) for Speaker 2. Tongue Reaction Time was on average 670 ms (SD 237 ms) for Speaker 1 and 442 ms (SD 137 ms) for Speaker 2. Observed Articulatory Reaction Time thus occurred approximately 175-180 ms ahead of Acoustic Reaction Time for Speaker 1. For Speaker 2 observed Articulatory Reaction Time occurred approximately 120-145 ms ahead of Acoustic Reaction Time. In other words, silent articulation comprised 20% - 25% of the Reaction Time as computed from the acoustic onset to speech (cf. Figure 2).

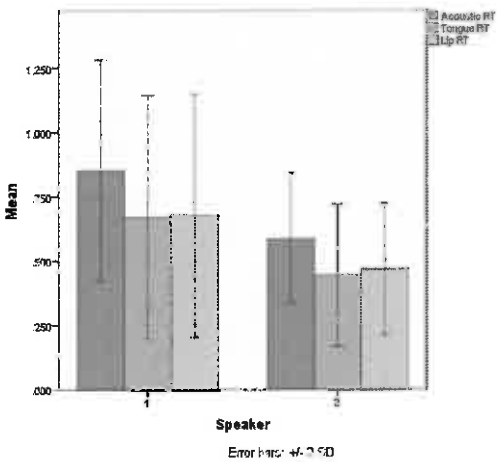


Figure 2: *Acoustic, Lip and Tongue RT (in seconds) per speaker.*

Within tokens, the mean difference between Lip and Tongue Reaction Time was 7 ms (SD 80 ms) for Speaker 1, which means there was no significant difference between the detection of Reaction Time using lip vs. tongue motion. For Speaker 2, the mean difference between Lip and Tongue Reaction Time within tokens was 24 ms, so Speaker 2's Lip Reaction Time was significantly slower than her Tongue Reaction Time (paired samples t-test, $t(76)=3.672$, $p<.01$).

4.3. Phonemic target types

Since each target lexeme was coded for the primary place, manner, and voicing of its initial consonant or vowel, we were able to test whether these phonemic targets gave rise to differently timed initiations of lip or tongue movement. There was no pattern that would suggest that one articulatory measure, lip or tongue, provided a more advantageous route to uncovering the Reaction Time for a certain phonemic target type. 'Onset Place', 'Onset Manner' and 'Number of Syllables' showed no significant effects on Reaction Time measurements.

Only 'Onset Voicing' showed a trend, with the trend going in opposite directions for the two speakers. For Speaker 1 Onset Voicing only had an effect on Acoustic Reaction Time (independent samples t-test, $t(123)=1.756$, $p=.082$), with Acoustic Reaction Times being on average 67 ms *shorter* for voiceless onsets than for voiced onsets. For Speaker 2 Onset Voicing showed a trend only for Tongue Reaction Time ($t(71.8)=1.769$, $p=.081$), with Tongue Reaction Times being on average 50 ms *longer* for voiceless onsets than for voiced onsets (cf. Figure 3).

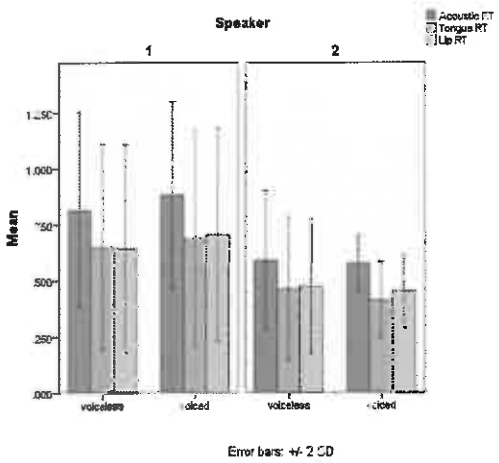


Figure 3: *Acoustic, Lip and Tongue RT (in seconds) per speaker and onset voicing status.*

This overall lack of phonemic bias is particularly interesting in light of findings that Reaction Times derived via the commonly used voice key are in fact quite susceptible to differences in onset type. Kessler et al. (2002) report striking phonetic biases, caused by phonemes' intrinsic differences in sound pressure level. Surveying data from four large-scale studies, they demonstrate that voiceless, posterior, and obstruent consonants set off voice keys later than others.

5. Discussion and conclusion

Overall, the findings demonstrate that articulatory measurements capture Verbal Reaction Time reliably at a much earlier time point than acoustic measurements, and while this is a small data set with only two speakers it raises interesting points related to inter-individual speaker differences. These are typical speakers, but also very different speakers.

Speaker 1 had not just fuller lips than Speaker 2, but also a tendency to move lips more, and maybe with a wider range. This might explain why Speaker 2 exhibited what seemed a slower Lip Reaction Time compared to her Tongue Reaction Time: We might not have been able to detect very early onsets of lip movement as easily as in Speaker 1, and lip movement might also just not be an as prominent feature in articulation in Speaker 2 as it is in Speaker 1. To estimate the impact of inter-individual (habitual or anatomical) differences on articulatory measures a much larger sample of speakers is required.

Another interesting point is the substantial difference in overall Reaction Time: Speaker 2 was overall much faster in

her verbal response than Speaker 1. Speaker 2's advantage of Articulatory over Acoustic Reaction Time was smaller but actually in proportion with her overall 'compressed' run-up to audible speech: For both speakers the duration of the Articulatory Reaction Time amounted to only around 75 - 80% of the Acoustic Reaction Time.

This suggests that there is a robust advantage of articulatory measurements of Reaction Time to measurements based on acoustic data – no matter how quickly speakers are able to get their articulators into motion. A systematic quantification and qualification of the articulatory data advantage could potentially inform a re-evaluation of well established Reaction Time protocols protocols based on analysis of facial video or Ultrasound Tongue Imaging.

6. References

Articulate Instruments Ltd (2008). *Ultrasound Stabilisation Headset Users Manual: Revision 1.4.* Edinburgh, UK: Articulate Instruments Ltd.

Articulate Instruments Ltd (2012). *Articulate Assistant Advanced User Guide: Version 2.14.* Edinburgh, UK: Articulate Instruments Ltd.

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. In: *Journal of Cognitive Neuroscience*, 22(7), pp. 1504-1529.

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. In: *Journal of Memory and Language*, 47, pp. 145–171.

Levelt, W., Roelofs, A. & Meyer, A. (1999) A theory of lexical access in speech production. In: *Behavioral and Brain Sciences*, 22, pp. 1–75.

Mooshammer C., Goldstein L., Nam H., McClure, S. & Saltzman E. (2012) Bridging planning and execution: Temporal planning of syliables. In: *Journal of Phonetics*. 40: pp. 347–389.

Palo, Pertti, Schaeffler, S.& Scobbie, J.M. (2014). Pre-Speech Tongue Movements Recorded with Ultrasound. In: *Proceedings of the 10th International Seminar on Speech Production*, Cologne.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. In: *Cognition*, 64, pp. 249-284.

Scobbie, J.M., Wrench, A.A., & van der Linden, M. (2008) Head-Probe Stabilisation in Ultrasound Tongue Imaging Using a Headset to Permit Natural Head Movement. In: *Proceedings of the 8th ISSP*, Strasbourg, pp. 373-376.

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. In: *Journal of Experimental Psychology: Human Learning and Memory*, 6, pp. 174-215.

Wrench, A.A. and Scobbie, J.M. (2011) Very high frame rate ultrasound tongue imaging. In: *Proceedings of the 9th International Seminar on Speech Production*, Montreal, pp. 155-162.

Wrench, A.A. and Scobbie, J.M. (2006) Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. In: *Proceedings of the 7th International Seminar on Speech Production*. Ubatuba, pp. 451-458.